学校编码：10384　　分类号_____密级_____
学号：20051302110　　UDC_____

厦门大学

硕 士 学 位 论 文

细胞因子-受体相互作用的预测

Genome－Wide Prediction of Cytokine-Receptor Interaction

许津瑞

指导教师姓名：纪志梁 副教授

专 业 名 称：生物化学与分子生物学

论文提交日期：2008 年 4 月 25 日

论文答辩时间：2008 年 6 月 3 日

学位授予日期：

答辩委员会主席：_____

评 阅 人：_____

2008 年 06 月

厦门大学学位论文原创性声明

兹呈交的学位论文，是本人在导师指导下独立完成的研究成果。本人在论文写作中参考的其他个人或集体的研究成果，均在文中以明确方式标明。本人依法享有和承担由此论文而产生的权利和责任。

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人完全了解厦门大学有关保留、使用学位论文的规定。厦门大学有权保留并向国家主管部门或其指定机构送交论文的纸质版和电子版，有权将学位论文用于非赢利目的的少量复制并允许论文进入学校图书馆被查阅，有权将学位论文的内容编入有关数据库进行检索，有权将学位论文的标题和摘要汇编出版。保密的学位论文在解密后适用本规定。

本学位论文属于

1、保密（　），在　　年解密后适用本授权书。

2、不保密（　）

（请在以上相应括号内打“√”）

作者签名：　　日期：　　年　月　日

导师签名：　　日期：　　年　月　日

目　录

CONTENTS

摘　要

细胞因子是由多种细胞所分泌的低分子量(15～30KD)的蛋白或糖蛋白，通过与相应受体的结合，调节细胞的生长和分化，参与免疫、炎症反应和创伤愈合。因为细胞因子通过与受体的相互作用来行使以上功能，所以细胞因子-受体相互作用成为研究细胞因子功能的关键。然而，由于细胞因子以网络形式行使生物功能，众多细胞因子在机体内相互促进或相互抑制。这一特性使我们对细胞因子的新功能和新细胞因子受体的存在所知甚少，又因为传统实验方法费用高，耗时长，所以应用计用算方法系统研究细胞因子-受体相互作用，以揭示细胞因子的新功能、新受体十分必要。本课题首先建立基于支持向量机的分类模型，用来从人类蛋白组中预测新的细胞因子-受体相互作用，最终得到 1609 个新相互作用，由 244 个新受体和所有 112 个已知受体参与。通过改进训练方法，该模型与常规支持向量机方法训练的模型相比有较高的准确性(99.2%)。进一步分析显示，各家族的细胞因子受体趋向于享有共同的序列模式(Domains/Motifs)，以便保证与细胞因子的特异结合。基因表达模式分析发现细胞因子与其相应受体的基因表达只有微弱的组织相关性。染色体分布研究发现人类基因组有 11 个明显的已知和/或预测细胞因子受体集中分布区。预测的相互作用参与了 31 个细胞 Pathway，其中 9 条为首次发现可能与细胞因子有关。

关键词: 细胞因子，细胞因子受体，支持向量机

ABSTRACT

Cytokines, 15～30KD, are a diverse group of cell intercellular messengers responsible for signaling variety of cell functions, such as immunity, hematopoiesis, chemotactic activities, cell maturation, proliferation, growth and differentiation through their interactions with respective receptors on cell membranes. The binding of cytokine to its complementary receptor is crucial for triggering cytokine-specific cellular activities. In this study, a series of Bioinformatics and statistical analyses were conducted to probe cytokine-specific activities by identifying cytokine-receptor interactions in the human genome. An advanced support vector machines system CytoSVM was constructed to successfully identify (estimated prediction accuracy 99.2%) 1,609 novel cytokine-receptor interactions from human genome. This covers 244 distinct novel cytokine receptors and 112 known cytokine receptors. To characterize the cytokine-receptor interactions, systematic and statistical analyses were demonstrated to illustrate their structural, transcriptional and genomic features. It is found that cytokine receptors intend to share some common signature patterns (domains/motifs) inside families. These patterns are conserved and specific in respective cytokine receptor families for execution and maintenance of their cytokine-receptor interactions. Data mining of the gene expression profiles revealed that only weak correlation between the expressions of cytokines and their complementary receptors. Chromosomal localization of cytokine receptors exhibits 11 obvious clusters of known and/or putative cytokine receptors distributed in human chromosomes; some of which are seem to be cellular activities-related. In particular, these newly identified interactions are involved in 31 characterized cellular activities in KEGG pathway database, 9 of these activities haven't been well studied or found for the known cytokines.

Key words: Cytokine, Cytokine Receptor, Support Vector Machine

1. 前言

细胞因子

细胞因子的定义

细胞因子是由多种细胞所分泌的，具有调节细胞的生长和分化、调节免疫、参与炎症发生和创伤愈合等功能的小分子多肽的总称[1] 。细胞因子的种类繁多，生物学作用各异，作为细胞间信号分子，它们在多种细胞活动中起着重要作用。

细胞因子的种类及功能

根据结构和主要功能，细胞因子可被分为以下几类[2] ：

白细胞介素(Interleukin，IL)是由淋巴细胞、单核细胞或其他非单核细胞产生的细胞因子，在细胞间相互作用、免疫调节、造血以及炎症过程中起重要调节作用。目前已经命名的白细胞介素有 IL1-IL23。

集落刺激因子(Colony-Stimulating Factor，CSF)根据其刺激造血干细胞或不同分化阶段的造血祖细胞在固体培养基中形成不同细胞集落的特性，分为粒细胞集落刺激因子(G-CSF)、巨噬细胞集落刺激因子(M-CSF)、粒细胞-巨噬细胞集落刺激因子(GM-CSF)和多重集落刺激因子(Multi-CSF，IL3)、干细胞因子(SCF)、红细胞生成素(Erythropoletin，EPO)、血小板生成素(Thrombopoietin，TPO)和 Flt3 配体(Flt3：FMS 样酪氨酸激酶 3)等。CSF 不仅可刺激不同发育阶段造血干细胞和祖细胞的增殖和分化，有的还可促进成熟细胞的功能。

干扰素(IFN)最初因发现病毒感染的细胞能产生一种物质可干扰另一种病毒的感染和复制，而将该物质命名为干扰素。根据来源和结构，干扰素分为 IFN-α、IFN-β、IFN-ω 和 IFN-γ，它们分别主要由白细胞、成纤维细胞和活化 T 细胞所产生。各种不同的 IFN 生物学活性基本相同，具有抗病毒、抗肿瘤和免疫调节等作用。

肿瘤坏死因子(Tumor Necrosis Factor, TNF)根据其来源和结构不同，可分为由单核/巨噬细胞产生的 TNF、由活化 T 细胞产生的淋巴毒素(Lymphotoxin)和由活化 T 细胞表达的膜型淋巴毒素。肿瘤坏死因子除具有杀伤肿瘤细胞外，还有

免疫调节、参与发热和炎症的发生等效用。

转化生长因子-β(Transforming Growth Factor-β,TGF-β)是属于一组调节细胞生长和分化的 TGF-β 超家族。这一家族除 TGF-β 外，还有活化素(Activins)、抑制素(Inhibins)、缪勒氏管抑制质(Mullerian Inhibitor Substance, MIS)和骨形成蛋白(Bone Morpho-genetic Proteins, BMPs)。TGF-β 的命名是根据这种细胞因子能使正常的成纤维细胞的表型发生转化，即在表皮生长因子(EGF)同时存在的条件下，改变成纤维细胞壁生长特性而获得在琼脂中生长的能力，并失去生长中密度信赖的抑制作用。

趋化因子家族(Chemokine)包括四个亚家族 CXC 亚家族、CC 亚家族、CX3C 亚家族、C 亚家族。趋化因子的主要作用是趋化细胞的迁移。有些趋化因子在免疫监视过程中控制免疫细胞趋化，如诱导淋巴细胞到淋巴结。 这些淋巴结中的趋化因子通过与这些组织中的抗原提呈细胞相互作用而监视病原体的入侵。有些趋化因子在发育中起作用；他们能刺激新血管形成；提供具体的关键信号而促成细胞成熟。有的趋化因子也可以促进伤口愈合。

其他细胞因子 如表皮生长因子(EGF)、血小板衍生生长因子(PDGF)、成纤维细胞生长因子(FGF)、肝细胞生长因子(HGF)、胰岛素样生长因子-1(IGF-1)、白血病抑制因子(LIF)、神经生长因子(NGF)、抑瘤素 M(OSM)和血管内皮细胞生长因子(VEGF)等。

细胞因子的特征

细胞因子的生物学特征　绝大多数细胞因子是低分子量(15～30KD)的蛋白或糖蛋白[2] 。天然的细胞因子由抗原、丝裂原或其他刺激物活化的细胞分泌。多数细胞因子以单体形式存在，少数细胞因子如 IL-10、IL-12、M-CSF、TGF-β、PDGF 等以双体形式存在，TNF 可形成三聚体。细胞因子通常以非特异方式发挥作用，即细胞因子对靶细胞作用无抗原特异性，也不受 MHC 限制。大多数细胞因子都以较高的亲和力与其受体结合，因此，很微量的细胞因子就可对靶细胞产生显著的生物学作用。

细胞因子的作用方式　细胞因子可以旁分泌(paracrine)、自分泌(autocrine)

或内分泌(endocrine)的方式发挥作用。若某种细胞因子的靶细胞（细胞因子作用的细胞）也是其产生细胞，则该因子对靶细胞表现出的生物学作用称为自分泌效应；若某种细胞因子的产生细胞和靶细胞非同一细胞，且二者邻近，则该因子对靶细胞表现出的生物学作用称为旁分泌效应。少数细胞因子如 TGF-β、IL-1 和 M-CSF 在高剂量时也作用于机体远处的靶细胞，表现内分泌效应。

细胞因子的网络性　一种细胞可产生多种细胞因子，不同类型的细胞也可产生一种或几种相同的细胞因子。一种细胞因子可对多种靶细胞发生作用，产生多种不同的生物学效应，这种性质称多效性；几种不同的细胞因子也可对同一种靶细胞发生作用，产生相同或相似的生物学效应，这种性质称为重叠性；一种细胞因子可以抑制另外一种细胞因子的某种生物学作用,表现为拮抗效应；可以增强另一种细胞因子的某种生物学作用，表现为协同效应。见图 1，众多细胞因子在机体内存在，相互促进或相互抑制，形成十分复杂的细胞因子网络[3] 。

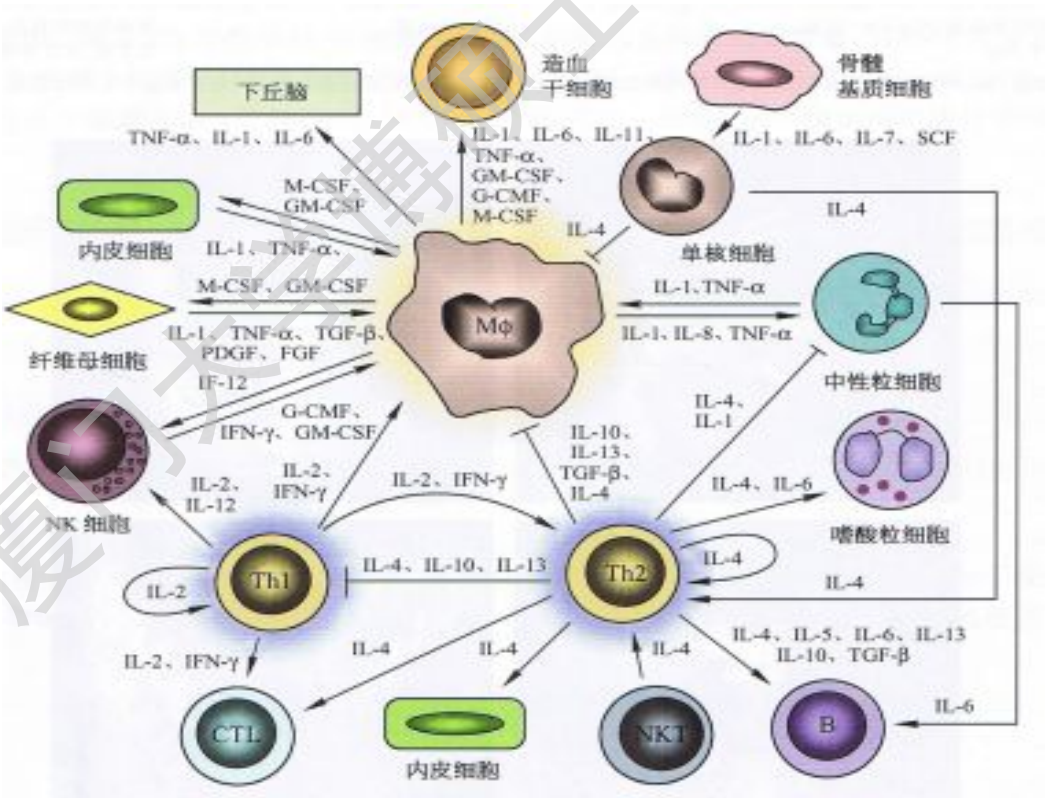

图 1. 细胞因子网络

Figure 1. Cytokine Network

目前，虽然细胞因子的功能还为完全阐明，但其功能均通过与受体的相互作用来实现，因此，研究细胞因子-受体相互作用对探究细胞因子的新功能至关重要。然而，由于细胞因子及其受体具有浓度小、半衰期短、多效性等特点，其分离纯化和功能研究较为困难。生物信息学方法研究细胞因子-受体相互作用能够克服以上问题，从而弥补实验方法的不足。

细胞因子-受体相互作用属于蛋白-蛋白相互作用(Protein-Protein Interaction)。这一课题已成为生物信息学研究的热点。

蛋白相互作用的研究方法

蛋白质相互作用的研究手段繁多，包括一系列已经建立的实验方法如酵母双杂交系统[4] 、质谱仪方法[5] 和蛋白质芯片[6] 等。近年来，随着计算机科学的发展，计算方法已经成为蛋白相互作用研究中一个有力工具。

预测蛋白质的相互作用是目前生物信息学中热门的研究领域，因为计算模拟的方法要比大部分的实验方法用时短，花费少。近几年，一些计算机科学中的算法已被用来预测蛋白质间的相互作用。总的说来，这些方法可以分成四类：1)基于基因组信息的方法；2)基于进化关系的方法；3)基于蛋白质序列的从头预测方法；4)需要三维结构信息的方法。基于基因组信息的预测方法包括系统发育谱(Phylogenetic Profile)[7] [8] 、基因邻接(Gene Neighborhood)[9] [10] 、基因融合(Gene Fusion Event)[11] [12] 以及镜像树(Mirror Tree)[13] 等方法。而基于进化信息的方法包括突变关联(Correlated Mutation)[14] 、保守的蛋白质相互作用(Interologs)[15] 、进化速率关联 Ccorrelated Evolutionary-Rate)[16] 等方法。上述方法均不可避免地具有一定的局限性，即它们都需要一些蛋白质的先验知识，如基因组信息、进化信息等。

而基于蛋白质的一级结构的预测方法首先由 Bock 和 Gough[17] 提出，该方法不需要基因组或进化的信息，仅仅需要单个蛋白质的序列信息。该方法从 DIP 数据库[18] 中提取相互作用的蛋白质的序列数据，根据蛋白质对的序列信息，包括氨基酸残基的理化特性、电荷以及疏水特性等，用支持向量机的方法训

练，其交叉验证的结果表明了该方法具有很高的准确率，大约在 80%左右。但他们的方法仅仅能鉴定真实蛋白质对和“假蛋白质”对，不能解决实际的问题。

最近，有人提出了利用蛋白质的三维结构信息进行蛋白质相互作用预测的方法，即同源结构复合物(Homologous Structural Complexes)方法[19] 。该方法构建了一种全新的策略，并提供了在线服务(http://www.russell.embl.de/interprets)，该数据库包括了 429 对非冗余相互作用结构域，1131 个已知三维结构的复合体，用提交的两个序列对数据库搜索序列的同源性，如果找到一个同源序列，即可证明该提交的蛋白质序列存在相互作用有潜能。另外，在一对随机序列背景的基础下，一种用来估计蛋白质潜在相互作用的具有统计意义的方法也已经被提出，Lichtarge 等[20] 据此设计了在进化中有意义的重要氨基酸的聚类方法，用来在三维空间里进行蛋白质的功能位点的预测。

机器学习方法及其支持向量机的基本原理

机器学习方法概述

基于数据的机器学习方法是现代智能技术中的重要方面，研究从观测数据样本出发寻找规律，利用这些规律对未来数据或无法观测的数据进行预测。机器学习的策略大体可分为：机械式学习、讲授式学习、演绎式学习、归纳式学习、解释学习和类比学习[23] 。

机器学习的实现方法大致可分为三种：第一种是经典的(参数)统计估计方法。包括模式识别、神经网络等。第二种方法是经验非线性方法，如人工神经网络(ANN)。第三种是基于统计学习理论的方法，如在此理论基础上发展的一种新的通用学习方法——支持向量机(Support Vector Machine，SVM)[24] 。

支持向量机理论简述

1．最优分类面

支持向量机方法建立在统计学习理论的 VC 维理论和 Vapnik 提出的结构风险最小化原理基础上，其主要针对两类分类问题，在高维空间中寻找一个超平面作为两类的分割，以保证最小的分类错误率。在二维情况下支持向量机的原理如

图 2 所示，其中的实心点和空心点代表两个类别的训练样本，H 为将这两个类别分开的分类线，H1 和 H2 分别是经过这两个类别样本中距分类线最近的点且平行于分类线的直线，H1 和 H2 之间的距离叫做这两个类别的分类间隙。支持向量机的目标是找到最优分类线，最优分类线不但能将两个类别的样本准确分开，而且要使两类的分类间隙最大。

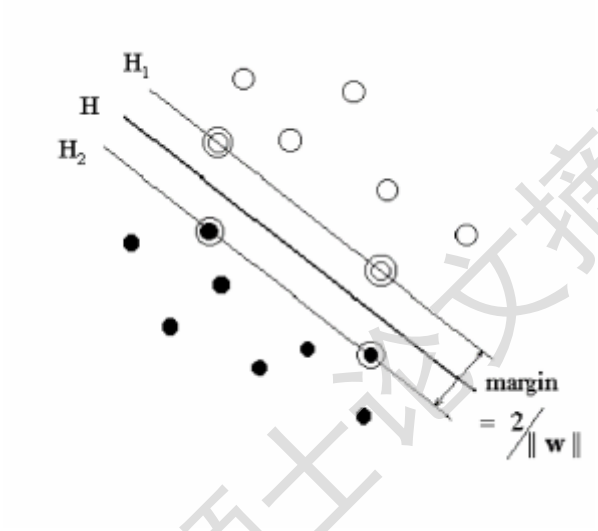

图 2 特征空间中的最优分割平面

Figure 2. Optimized hyperplane in future space

如图 2，考虑一个用某特征空间的超平面对给定训练数据集做二值分类的问题。对于给定样本点：

$$(x_1, y_1), \mathrm{K}, (x_l, y_l), x_i \in R^n, y_i \in \{-1, +1\} \tag{1}$$

设分类线为：

$$y(w \cdot x + b) - 1 = 0 \tag{2}$$

其中 w 为分界线的权系数，b 是分类阈值。则间隔为 $\frac{1}{2}\|w\|$，

则间隔最大等价于使 $\frac{1}{2}\|w\|^2 \to \min$，约束条件：

$$y_i[(w \cdot x_i) + b] \geq 1, i = 1, 2, ..., \tag{3}$$

满足上述条件的分类线为最优分类线，此时 H1，H2 上的训练样本点称作支持向量，因其支撑了最优分类线(面)。

在线性可分情形下构造最优超平面是利用二次规划求解最优问题，有唯一极小点，用 Lagrange 乘子法把式(3)化成对偶形式

$$\sum_i \alpha_i - \frac{1}{2}\sum_{i,j} a_i a_j y_i y_j (x_i . x_j) \to \max$$
$$\text{约束条件：} \alpha_i \geq 0, \sum_i a_i y_i = 0 \text{ , i = 1,2,...,} \tag{4}$$

$$w(\alpha) = \sum_{i=1}^{n} \alpha_i y_i x_i \tag{5}$$

即最优分类面的权系数向量是训练样本向量的线性组合。且根据 Kuhn-Tucker 条件，该最优化问题的解须满足：

$$a_i (y_i (w.x_i) + b - 1) = 0, i = 1,2,..., \tag{6}$$

因此，对多数样本 ai 将为零，取值为零时 ai 对应于使式(3)等号成立的样本即支持向量，它们通常只是全体样本中的一小部分。解上述问题后得到最优分类函数：

$$f(x) = \text{sgn}\{(w.x) + b\} = sign\{\sum_i a_i y_i (x_i . x + b)\} \tag{7}$$

式中的求和实际上只对支持向量进行，b 可以用任一个支持向量(满足式(3)中的等号)求得，或通过两类中任意一对支持向量取中值求得。

2．广义最优分类面

若训练样本是线性不可分的，或事先不知道它是否线性可分，则与式(2)变为

$$y_i (w \cdot x_i + b) \geq 1 - \xi_i, \xi_i \geq 0, i = 1,..., n \tag{8}$$

与式(3)相应的优化问题是： $Min(\frac{1}{2}\|w\|^2 + C\sum_i \xi_i)$ (9)

这里 ξ i 可看作训练样本关于广义分离超平面的偏差， ξ i= 0 时问题变为线性可分情形，c> 0 是自定义的惩罚系数，用来控制样本偏差与机器泛化能力之间的平衡。用 Lagrange 乘子法把式(9)化成其对偶形式，其结果与线性可分情况下几乎完全相同，只是约束条件式(4)变为： $0 \leq a_i, \leq c, i = 1,..., n$ (10)

3．支持向量机

超平面分类能力有限，为此引入分类曲面，主要思想是作非线性映射